We need an American DeepSeek.

We need a sovereign, *open source* AGI powerhouse and we need it now.

That's because American frontier labs are losing the race.

Don't believe me?

For every one American closed or proprietary model, we're seeing dozens of industrial grade open models out of the Middle Kingdom.

The only American open source champion, Meta, put out the last Llama model, Llama 3, a year ago on April 2024. It's since been swept away by a tsunami of Chinese models. A year is a lifetime in AI. And even worse, its subpar multimodal capabilities don't hold a candle to the multimodal power of Chinese models.

Llama 4 will likely deliver amazing results but one American powerhouse can't keep up with the breakneck speed of releases needed to compete against 4 or 5 or 10 open source AI companies out of China dropping models every month.

Since Llama 3's initial release we've had DeepSeek V2 and V3 (and a new V3 iteration just this week), as well as the shockingly capable DeepSeek R1, along with Qwen 2 and 2.5, powerful text only models, and Qwen 2.5 VL (probably the most powerful multimodal model available right now outside of the closed alternatives, trained on everything from video, to GUI navigation, to agentic tasks), with updates like the 32B model just dropping, along with models from ByteDance and more, not to mention the absolute storm of video models out of China that have absolutely crushed Sora.

The new LLama is coming fast but the rumor mill was churning after the release of DeepSeek R1 that Meta engineers had scrambled a war room to tear apart its papers trying to figure out where they went wrong with Llama 4 and it's since been delayed while they catch up to R1's reasoning capabilities.

Think American chip sanctions will keep the Chinese AI juggernaut at bay? They won't. China is pouring billions and billions into creating cheaper, faster, better EUV light machines and they're testing them right now. Almost certainlty, the chip sanctions will be utterly worthless in just a few years time. All they'll do is guarantee the rest of the world is buying Chinese chips instead of American ones.

Something has got to change and it's got to change fast.

We need more American open source champions and we need them now.

Even worse, not only are American labs not releasing open source, sometimes they're not releasing at all.

Sora came out to great fanfare on X but then didn't come out for almost a year later, when it had already been outclassed by other video models at home and abroad. While American labs have been infested by EA driven "safety" pogroms and obsession with centralized control, Chinese companies just keep releasing at a furious pace and laughing as we shoot ourselves in the foot.

There are multiple reasons for American frontier lab stumbles but the biggest is capacity. GPUs are massively expensive and nobody can buy or rent enough to meet growing demand. OpenAI dropped GPT 4o image gen and a few days later Altman tweeted their GPUs were "melting." Less than a week after the release of Gemini 2.5 the head of Google AI Studio said they were still "constrained by rate limits" giving developers a measly 20 requests per minute. No single entity can seemingly keep up with the inference demands of mega-models.

Sometimes the reasons are commercial. o3 made massive waves by smashing the first iteration of ARC-AGI but where is it? Supposedly it powers DeepResearch but nobody else outside of OpenAI can get their hands on it, just smaller distilled versions like o3-mini. They'd rather it power their tools than let you get your hands on it to build a competitor to DeepResearch.

Other times it's just the size of these models, which points back to capacity. Newer models like 4.5 are huge, trained on 10X more compute, and yet they're prohibitively expensive and basically no better than their older brothers. The pre-training scaling laws are dead.

Sometimes it the culture of anti-tech and degrowth that’s poisoning American innovation and its historical optimism. The big American labs have been positively infected with people who want to stop, slow down or cripple AI with chains and guardrails or just stop it from happening at all.

LOOK EVERYONE. I JUST INVENTED A WHEEL. JUST THINK OF ALL THE GREAT USES. WE'LL BE ABLE TO MOVE THINGS FROM ONE PLACE TO ANOTHER. IT WILL HELP WITH CONSTRUCTION AND ALL SORTS OF THINGS WE CAN ONLY DREAM OF!
NO BIG DEAL. IT CAN'T EVEN AUTOSTEER!
SCALING IT UP WILL NEVER ACHIEVE SPEEDS OF 60mph!
BAD PEOPLE WILL USE IT TO DO BAD THINGS!
ONLY THE RICH WILL HAVE WHEELS!
WHEELS WILL BE USED FOR WAR!
WHAT COULD GO WRONG?
IT WILL TAKE THE JOBS OF PEOPLE WHO CARRY THINGS!

(Source: Neil Christensen and GPT-4o)

Meanwhile, Chinese AI powerhouses like DeepSeek, ByteDance and Alibaba have ripped a page from that old American playbook of global software dominance:

Win with open. Win with soft power. Win with good enough.

Is 10 points of difference on a benchmark worth 100x the cost to you?

Nope.

Not for 80% of the use cases.

And while American companies struggle to meet demand, Chinese companies are releasing like crazy.

If you release an open model you get the benefits of distributed scale. People can run it on their own GPUs. Since nobody can get enough GPUs to power all that demand, better to just release it and let the power of distributed GPUs power it. Yeah you don't capture all those dollars yourself, but you capture mindshare. And some of those dollars come back to you. When someone builds an enterprise app on an open model, they're likely to come back to you for faster enterprise inference and Max versions of the model to power that app in production, so its a virtuous circle.

Open is the winning strategy now and it will be the winning strategy tomorrow.

Over the coming decade, open models will form the bedrock of smart economy and we'll see a wild array of closed and open apps built on top of them.

Remember, open doesn't mean no closed apps. It means we commoditize one layer of the stack and then move up the stack to build more interesting things. Twitter is closed but it couldn't exist without open, commodity building blocks like Docker, Kubernetes, Linux and Pytorch.

But how does that even work? How do you give away software and make any money at all?

While open source companies don't profit from every deployment, the playbook for open source has always been apps built with commodity bricks, enterprise support and services for software, which is why companies like Red Hat, Hashicorp and MongoDB became multibillion dollar businesses while "giving away" their core products.

When it comes to AI that model will play out again.

Open source companies will commoditize intelligence and we'll move up the smart software stack.

We'll see enterprise services like "long context" versions of the models, finetunings for specific use cases, inference scaling, memory, apps, APIs, SDKs, and consulting to drive revenue around those models. We'll also see the open core model take root, where some parts of something like a Mixture of Infinite Experts model (where fresh experts or adapters are added to meet a specific use case) are open sourced while skillpacks of advanced capabilities are monetized. We'll also see sleek and integrated interfaces that people pay for and we'll see platforms for tuning and training the models even more.

DeepSeek was the turning point for AI. It's taken off inside and outside of China, showing rapid adoption at companies like Perplexity, Amazon and enterprises everywhere.

Over the last two decades American-led open source swept across the globe, creating trillions in economic value, powering everything from the phone in your pocket, to the router in your house, to American mega-clouds like AWS, Azure and GCP, to 100% of the most powerful supercomputers on Earth.

On a pound for pound basis, nothing creates value like open source at the civilization level. As Clem Delangue of Huggingface wrote recently about the Harvard study on open source:

> "$4.15B invested in open-source generates $8.8T of value for companies (aka $1 invested in open-source = $2,000 of value created) - Companies would need to spend 3.5 times more on software than they currently do if OSS did not exist."

Chinese companies have learned from it. American AI labs have forgotten it.

American labs have gone totally closed. They benefitted immensely from open research and open designs like the Transformer and Pytorch and Linux and Docker containers and Ray and yet they've gone in the exact opposite direction.

While American labs focused on closed source models while trying to make competition illegal by banning Chinese models from the American market and pushing for draconian chip export controls, Chinese labs keep winning with open.

But with all that said even open is not enough to win the race to AGI.

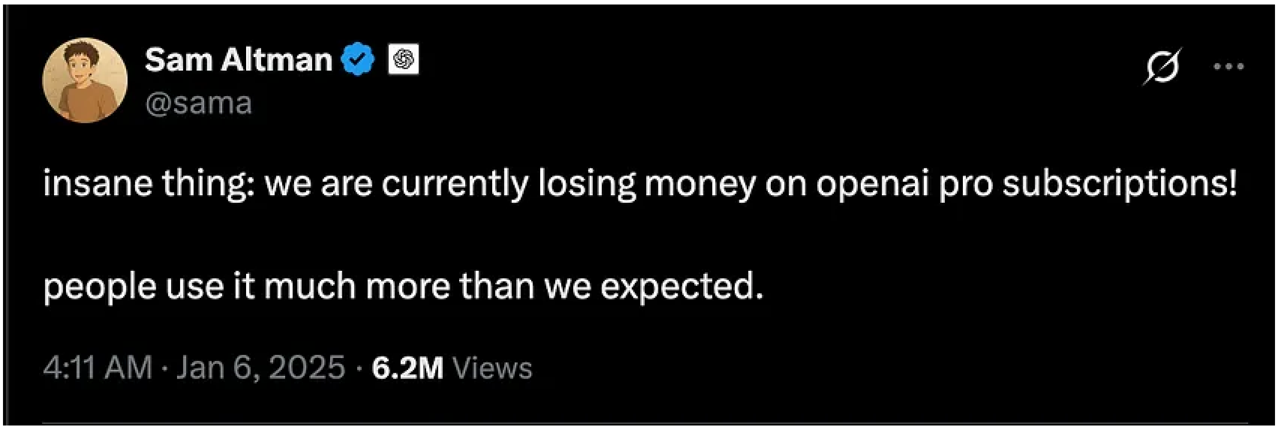
We need something different. Something more. We need to solve the biggest problem in AI.

And what's that?

How to make AI profitable.

Not only are our labs not profitable, they're bleeding money. Anthropic jumped into the Google anti-trust case because the government was looking to block the search engine giant's ability to invest in AI companies and Anthropic basically told the government they

needed the money to keep running. Despite a $200 pro subscription for their elite models, Altman said OpenAI is losing money on it. That's no surprise when a single new cutting edge GPU costs 30-50K a pop.



Despite having billions in investment and scores of eager customers around the world, they aren't turning a profit.

This can't last.

We need a brand new business model in AI, one that nobody's figured out yet. And what kind of business model?

One that solves the hardware problem.

What's the hardware problem? Only the hardest problem in AI right now. It's not new algorithms or how to make a better, smarter, faster model.

It's how to deal with the fact that GPUs are freaking expensive and they're nothing but a massive cost sink to our AI labs. They're a catch 22. They're essential and yet they're the reason labs keep losing money.

Companies raise and spend billions on fleets of GPUs that depreciate faster than a brand new car. Software companies need them and yet they're nothing but a nasty drag on the bottom line of every single AI company on the planet.

So far, in this early era of AI, the only person getting rich is Jensen.

So how do you solve *the* Catch 22 of the AI era?

We take a page from another playbook.

Amazon.

The AWS of AI

To solve the problem of AI hardware, you've got to flip that hardware from a cost sink into a profit center.

To do that, tomorrow's frontier AI labs need to transform from software only companies into vertically integrated software *and* hardware companies.

One way to do that is to adopt the AWS model.

So what is the AWS model?

Get incredibly good at building datacenters and building a platform around those services and then rent out that platform to create the "operating system" of the Internet. That was the original idea behind the first cloud service.

AI will become the intelligence layer of the Internet. Think ambient intelligence: a billion thinking machines plugged into billions of smart apps and services, threaded through every aspect of our lives, from the supply chain, to manufacturing, to defense, to entertainment, marketing and more.

There's a persistent myth that AWS got started because Amazon had a lot of excess capacity hanging around. It's wrong. The real reason AWS got started was that Amazon was a low margin business and because of that they had to get really good at building data centers and running them profitably.

> "As the team worked, Jassy recalled, they realized they had also become quite good at running infrastructure services like compute, storage and database (due to those previously articulated internal requirements). What’s more, they had become highly skilled at running reliable, scalable, cost-effective data centers out of need. As a low-margin business like Amazon, they had to be as lean and efficient as possible," writes Ron Miller in TechCruch, about How AWS Came to Be.

They also wanted to fix their tech debt. They'd grown massively and wanted to untangle the mess of distributed software they'd build as they grew into a clean, API driven, consumption architecture.

> "At that point, the company took its first step toward building the AWS business by untangling that mess into a set of well-documented APIs. While it drove the smoother development of Merchant.com (a now defunct platform), it also serve the internal developer...too, and it set the stage for a much more organized and disciplined way of developing tools internally going forward.
>
> “We expected all the teams internally from that point on to build in a decoupled, API-access fashion, and then all of the internal teams inside of Amazon expected to be able to consume their peer internal development team services in that way. So very quietly around 2000, we became a services company with really no fanfare,” said former AWS CEO (now Amazon CEO), Andrew Jassy.

For AI companies to get profitable they're going to have to do the same. They'll need to rent out GPU capacity at scale, while creating a slick, hyper-valuable software layer over the top of it all. AI isn't a low margin business, it's a negative margin business right now.

Think CoreWeave meets OpenAI, if OpenAI were truly open.

The strategy is to buy mega-clusters of GPUs, then reserve half of those clusters for training but rent out as much as 80% of the reserve when the lab isn't training or doing experiments, both as reserved instances and spot instances.

That makes a truly cutting edge frontier lab that can pay its own bills in short order and has a constantly monetizable supply line of advanced chips both for training and for a hardcore on-demand HPC business. It flips a GPU into an asset instead of a CapEx and OpEx monstrosity. More demand means these labs can buy more compute with debt financing. If you've got enough customers and more clamoring to get on board the banks will give you money to go buy more chips to deploy.

And it can't come fast enough because a brutal reality for America is looming fast.

The reality is we're not ahead. We're falling badly behind it and it won't be long before we're too far behind and Chinese models have become the standard for the world.

While American labs will still be arguing about how the government can save them from competition with chip controls and bans, China will have already developed their own EUV machines (it's happening much faster than American analysts think [they're still thinking 5-7 years and they’re going to be dead wrong] and they're going to get blindsided by how fast) and they'll be using their superior manufacturing at scale capacity to crank out a blistering array of high end chips and they'll flood the market with chips and models and intelligence that the world will lap up because its cheaper and good enough and eventually state of the art.

Every country we slighted by putting them on the tier II or tier III list for export controls will be avid buyers of exclusively Chinese chips.

(Source: X)

You see, the world doesn't care if the future of models is Chinese or American.

Only Americans care about that. The world will use whatever works and works well. If it's American, great. If it's Chinese, no problem. If it's Ethiopian, all good.

So if Americans want to win there's only one path forward:

Compete.

Either we build the better mouse trap or we watch as others build it from the sidelines.

There's still time.

But it's running out.

The Brutal Reality

The brutal reality is that America is losing in both AI and robotics. We're losing in the digital world and in the physical one.

It may seem like we're winning with AI but we're not even close.

Right now, Chinese companies are releasing a storm of highly capable models that are almost as good as their American closed counterparts at a fraction of the price. In some cases those models are even better than the closed version, most notably on agentic capabilities like navigating GUIs.

If there are no American alternatives or those American alternatives are a year old, American companies will build with Chinese models. It's already happening now. The secret weapon powering most American agentic companies right now are Chinese models like Qwen.

That's because Chinese models are highly tunable and controllable, unlike their closed alternatives. When DeepSeek R1 rocked the world on its release, a gaggle of people who are absolutely clueless about machine learning shouted to the high heavens that the Chinese models are censored. Just ask them about Tibet they screamed!

They were wrong.

Perplexity (and anyone who knows the 101 basics about ML) knows that you can remove that kind of bias in days with a little fine tuning, making them more tunable than their American counterparts. Perplexity dropped an uncensored R1 on HF in short order and it now powers the reasoning of their search engine.

> Today we're open-sourcing **R1 1776**, a version of the DeepSeek-R1 model that has been post-trained to provide unbiased, accurate, and factual information. Download the model weights on our HuggingFace Repo or consider using the model via our Sonar API.

(Source: Perplexity blog announcement on DeepSeek)

Try to finetune Claude or GPT to change its biases and watch how long it takes or how impossible it is because of their layers and layers of "safety" and censorship controls. Even if you manage to fine tune it, you'll probably run into their traditional software layers that try to protect against certain use cases but end up hitting you with false positives instead.

"Safety" is a tax on downstream systems.

In the early days of o1, the filter layers at open AI often falsely flagged my questions about my own personal machine learning experiments on reasoning, probably because these dumb safety layers saw my questions as an attempt to subvert their rules that hide their internal reasoning token traces despite me and my team doing nothing of the sort. That makes for a bad user experience. Luckily, they seemed to have learned from their mistakes and the behavior went away, but it doesn't mean I won't see false positives in other scenarios and it's a bitter reminder that if you're playing by someone else's rules they can rip anything away from you and shut you off over night.

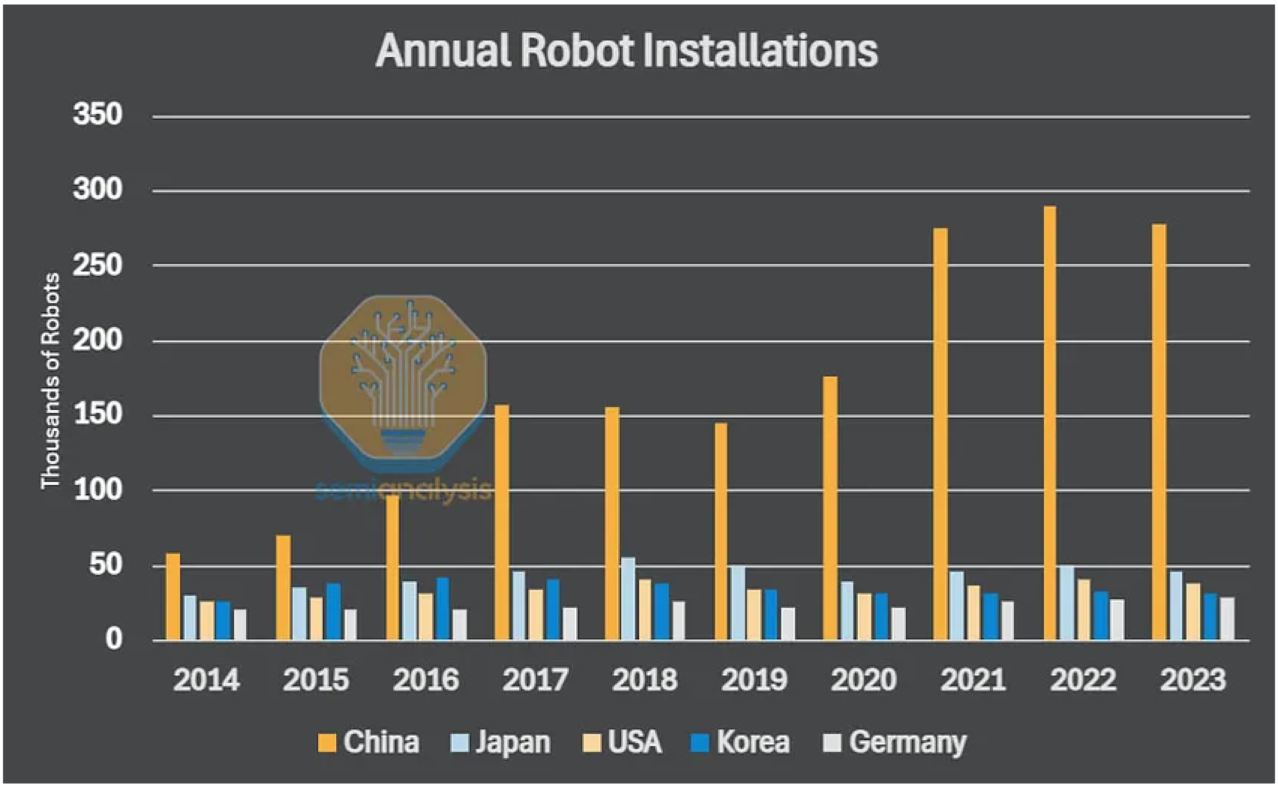
The recent GPT 4o image release is a perfect example of what happens when a few companies control what you can and can't do with your models. The first few days people where happily releasing a flurry of Ghibli inspired images. But a few days later they were already censoring Ghibli generations out of fear of a lawsuit.

As more and more Chinese models come out, more and more companies around the world will build with them, because they're cheaper and more tunable. They're not just good enough either. They're often better. Those models will gain mindshare and economic might. A robust ecosystem of fine tunes and additional capabilities will spring up around these models and before American companies know what hit them it will be too late.

When it comes to robotics the news for America is even worse off than that. Robotics and AI are closely linked. If models are the brains, then robots are their bodies in the real world.

In a brutal wake up call to the west called America is Missing the New Labor Economy, SemiAnalysis shows that America and the EU are badly behind on roboticization and the race to general purpose robotics.

When it comes to the sheer number of robots deployed in Chinese factories, nobody else in the world is even close. Take a look at that chart below. Even an idiot can see who is winning that race.



(Source: SemiAnalysis)

The authors paint a dire picture for the west:

> "This is a Call for Action for the United States of America and the West. We are in the early precipice of a nonlinear transformation in industrial society, but the bedrock the US is standing on is shaky. Automation and robotics is currently undergoing a revolution that will enable full-scale automation of all manufacturing and mission-critical industries. These intelligent robotics systems will be the first ever additional industrial piece that is not supplemental but fully additive - 24/7 labor with higher throughput than any human—, allowing for massive expansion in production capacities past adding another human unit of work.
>
> "The only country that is positioned to capture this level of automation is currently China, and should China achieve it without the US following suit, the production expansion will be granted only to China, posing an existential threat to the US as it is outcompeted in all capacities.
>
> "This is the manufacturing playing field that China has dominated for years now. The country has one of the most competitive economies in the world internally, where they will naturally achieve economies of scale and have shown themselves to be one of most skilled in high-volume manufacturing, at the same time their engineering quality has grown to be competitive in several critical industries at the highest level. This has already happened in batteries, solar, and is well underway in EVs. With these economies of scale, they are able to supply large developing markets, like Southeast Asia, Latin America, and others, allowing them to extend their advantage and influence."

Robotics is about to explode with a force that dwarfs past industrial shifts. Picture factories churning out robots that build still more robots. Each iteration cuts costs, boosts quality, and powers a production flywheel that keeps spinning ad infinitum—leaving rivals more and more desperate and breathless as they sprint to catch up but find out its hopeless.

Because robotics is a true general-purpose technology, it will upend everything from textiles and electronics to consumer goods.

Meanwhile, the West is about to get caught flat footed as the unexpected punch smashes into our face. Here's a fully automated factory in China cranking out the components for a thousand missiles a day 24/7. Meanwhile American factories are still producing missiles like its 1945, by throwing people at it. A Pentagon war simulation said the US runs out of missiles "in days" in a war with China because it can't match China's production capacity. Europe's industrial core is getting devoured by China's relentless rise and its absolute inability to source enough power (while shutting down their nuclear reactors), and the US keeps fixating on other markets while banking on cheap offshore production.

(Source: Weibo, "dark" factory in China)

All the while, China's manufacturing might gets stronger, its self-reliance grows faster, and robotics and intelligence are entering a bold new age.

This is not the recipe for the west to win and if we're not careful the next century is the century of the Middle Kingdom.

Even worse, we're asleep at the wheel. Americans think we're ahead with AI and robots and everything else. We imagine it's our birthright to be number one forever and forget that empires are born through a combination of constant innovation, new technology, creativity, acceleration, trade, business and brilliance.

That's how empires fail.

If we're not careful, we'll suffer the fate of every empire before us.

And that's because we've stopped seeing reality and we just see what we want to see. We're living in a daydream of our glory days in our heads. When the alarm rings on the new world, it's going to be a nasty surprise.

But all is not lost. The game is still being played and there's still time to turn it around.

To do it, we've got to wake up and do all of this and more:

- **Greenlight bold new nuclear energy initiatives at breakneck speed**

  (China has 150 new reactors planned by 2035)

- **Deliver massive investments in critical mineral mining wherever those deposits are in the world so we can make chips and everything else electronic**

  (Where critical minerals are)

- **Invest heavily in a public-private hybrid for advanced chip making capabilities, following the original TSMC playbook where the government clears roadblocks, lowers taxes for the chip giants and greenlights new energy sources**

  (History of TSMC)

- **Ramp up investments in next-gen robotics**
- **Build half a dozen open source intelligence powerhouses**

This is not a job for any single company. Nobody can do it alone.

We don't need more closed AI companies. We need a dozen vertically integrated ones that can massively scale hardware and software to dizzying new heights.

We need automated factories by the hundreds.

We need robots, lots and lots of robots.

There's some hope for American robotics already, some bright lights in the darkness. We've seen a huge investment by venture capital into humanoid robot companies like Figure and humans are good design. They may just be the design that gets us to universal robotics. Robots so far have been highly specialized designs that do one thing and do it well. But those robots are brittle and they can't adapt. If the riveting sheet is off by a small fraction of an inch the robot riveter arm fails. Humanoid and human spin off designs hold the possibility of adaptability.

But when it comes to open source AI labs in America we've got Meta and nobody else. That's not enough to stand against the tide of Chinese models coming out with blistering speed.

Even worse, our companies can't seem to make a profit as they choke down 30 or 40K GPUs that eat profits like a great white shark. We need profitable open labs.

One company can't do all of that and nobody should even hope to try, because it's a fools errand.

But it has to start somewhere. And the best place to start is a lighthouse American open intelligence lab, a brilliant light on the hill that inspires other American companies to follow suit, one that's cracked the hardware problem and makes amazing software too.

Do that and we might just turn it around.

Let's make AI open again.

While we're at it, lets solve the profitability problem for American's labs.

And if we do that, we just might see the American dream reborn for another day.

Future History is a reader-supported publication. To receive new posts and support my work, consider becoming a free or paid subscriber.

Discussion about this post

No posts